TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language

Jack Bowers [1,2,3]

iljackb@gmail.com

https://github.com/iljackb/Mixtepec_Mixtec

Laurent Romary [2,3,4]

[1] Austrian Center for Digital Humanities (ACDH)

[2] Inria – Team ALMAnaCH

[3] École Pratique des Hauts Études – Paris

[4] Berlin Brandenburgische Akademie der Wissenschaften (BBAW)

ICLDC 6 UH Mānoa – 2019/03/02

Mixtepec-Mixtec (Sa'an Savi)

- *Sa'an Savi* 'rain language'
- ISO 639-3 code: 'mix'
- Oto-Manguean
- San Juan de Mixtepec, Juxtlahuaca district (Oaxaca, MEX)
- Estimated 9-10,000 speakers; +~3000 in California, Oregon, Washington, Arkansas
- Status "vigorous" (Ethnologue 21st edition) but "threatened"(ELDP via Ethnologue 17th edition)
- Has been studied by: Pike and Ibach (1978); Paster and Beam de Azcona (2004-2007); Beckman and Nieves-SIL (2005-current)

Images: Mixtepec Mixtec | SIL Mexico. (n.d.). Retrieved March 2, 2019, from http://www.mexico.sil.org/language_culture/mixtec/mixtec-mix

Primary Sources of Language Data

- Consultation with Speakers
- Recordings made by speakers of others
- ~500 recordings, ~10hrs
- ~ 40 Booklets by SIL Mexico
- Public sources (Facebook, YouTube, etc.)
- Academic papers on language (5)
- Pamphlets from Mexican Government
- Bible (1 book) by Scripture Earth (+1hr spoken; 34,000 words)
- Personal Communication (texts, emails, facebook messages, etc.)

Desired Outcomes

- **Create an open source collection** of reusable and extensible LR
- Produce **corpus-based descriptions** and analyses of various aspects language's features to **further the knowledge of all aspects of the language**
- **Demonstrate and evaluate the application of encoding and standards** (particularly TEI) on an under-resourced non-Indo-European language
- Collect enough data so that it can be re-used for creating learning material

Specific Output

- Searchable (TEI) corpus from text- and time-aligned spoken sources
- Multilingual Digital (TEI) Dictionary (Mixtec, English, Spanish)
- Annotated (TEI) files of SIL booklets
- Lexical feature inventory
- Phonetic feature inventory
- Open Archive of all media and other files (Dataverse, Olac)
- Concepts inventory
- Place inventory
- Person list

Challenges Intrinsic to Studying Mixtepec-Mixtec

- Lack of existing resources (**under-resourced**)
- Need to gather from many **heterogenous sources**
- **Consultants limited availability** to edit, gloss text, etc.
- SIL Researchers working on the language in Mexico have (mostly) not shared their data (*though I do have agreement to reuse publications*)
- **Orthography** (in development by SIL) **not fully conventionalized**, still changes, speakers often not aware of/don't use the standards *(requires significant normalization in markup*)
- **Lexical tone, adds complexity** to characterization and it is (mostly) not represented in the orthography (*lot's of homographs*), creates irregularity in transcriptions
- **Not enough data** to automate annotation!

Overview of the Text Encoding Initiative (TEI)

- XML vocabulary began in 1987
- **Digital Humanities de facto standard** (open source consortium), used in lexicography, library sciences, digitizing archives, linguistic corpora, time-aligned speech
- **Dictionary well developed** (and increasingly widely adopted)
- **Pre-existing tag vocabulary, schemas, documentation**, (w/tools Oxygen XML editor) automatic validation, content completion, templates
- **Enables common corpus structure** for all levels of language (makes searching and retrieving easier)
- **Indepth metadata** available in header
- No limits on number of annotation categories
- Only need to document practices (not likely unique)
- **Freely available documentation**, mailing list, stable community with journal (jTEI), growing active users
- **Evolving to fit needs of users**, can submit issues, proposals to change (via GitHub)
- **Easy to generate human readable formats** (html, and/or using CSS)
- **Not software or platform dependant** (data-centered not platform-centered)

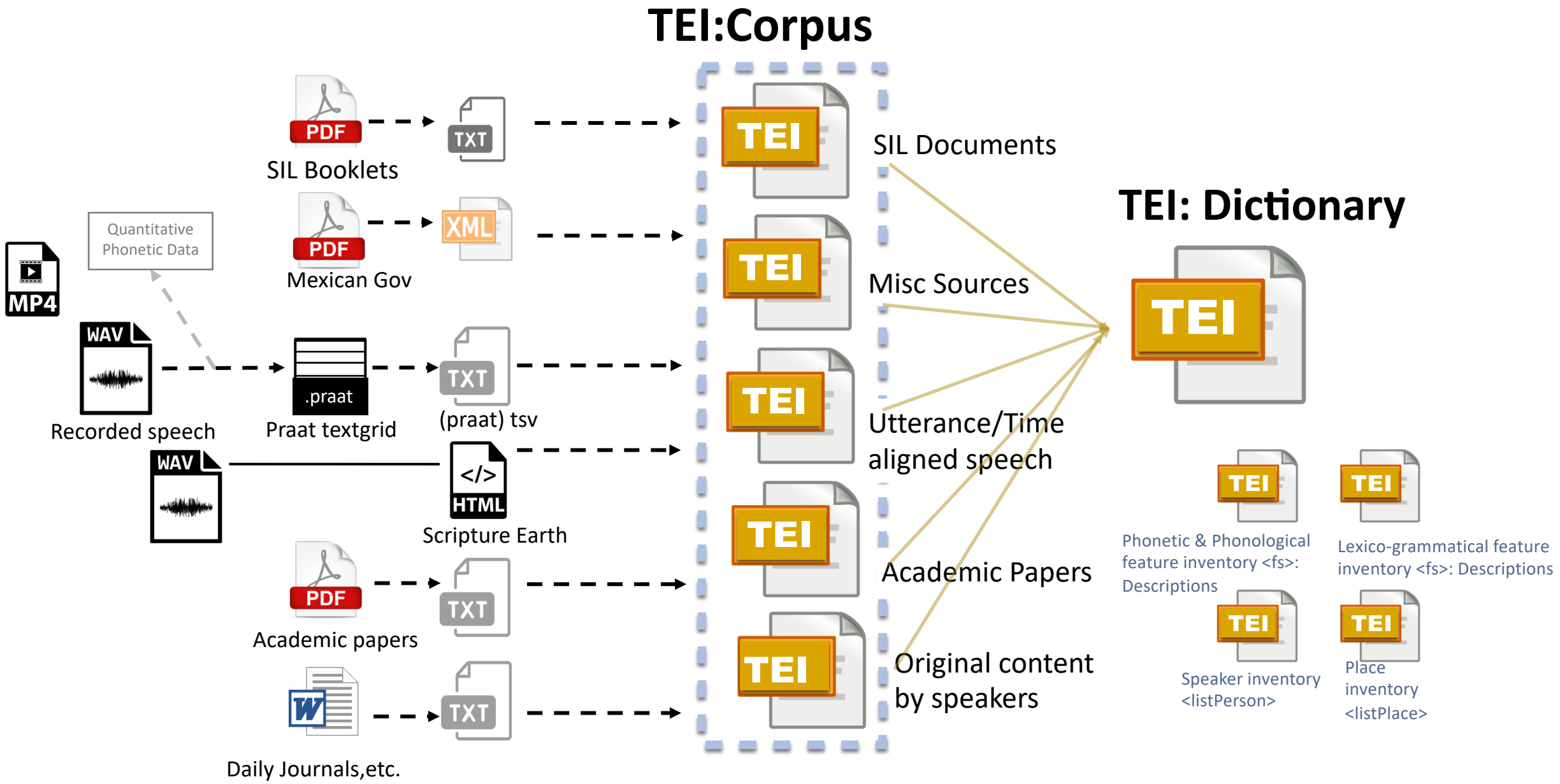
TEI for LD and Indigenous Languages

- **Underdeveloped** usage of standard
- **Mostly European languages**, increasing in Asia (esp. Japan) (only Czaykowska-Higgins et al. 2014 & Bowers and Romary 2018)
- **Already has appropriate components to handle all major aspects of LD project** (however nobody has previously used all), thus refinement needed
- **Enables integrated treatment** of variants, normalization, conflicting description, speaker attribution, etc.
- **Allows customization**

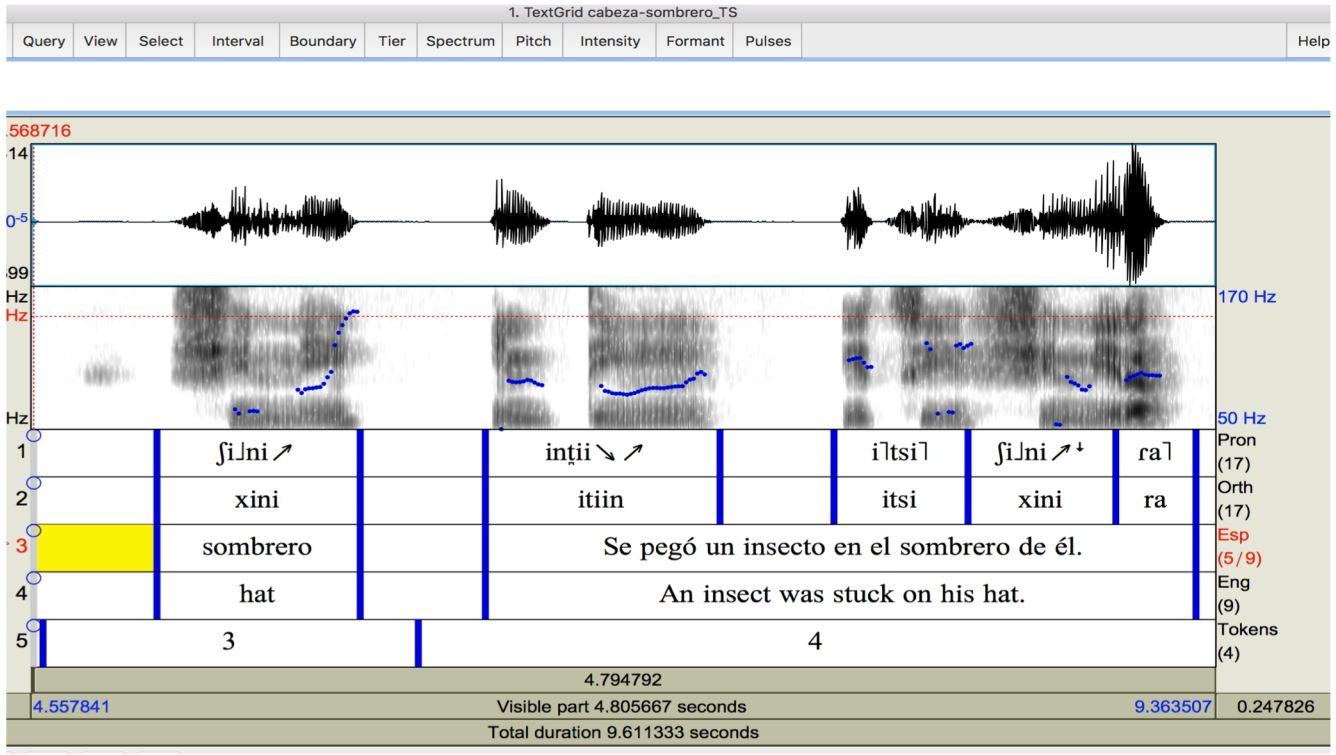
Czaykowska-Higgins, E., Holmes, M. D., & Kell, S. M. (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation and Conservation*, *8*, 1–37.

Bowers, J., & Romary, L. (2018). Bridging the gaps between digital humanities, lexicography and linguistics: a TEI dictionary for the documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America*, *39*(2).

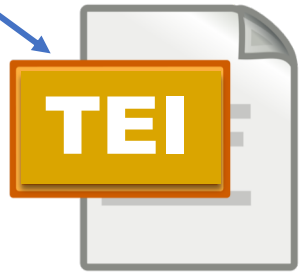
Mixtec Data: Sources, Workflow, Output



Speech annotation: Praat > TEI



tmin	tier	text	tmax
3.24	Pron	iˈtsiˈ	3.64
3.24	Orth	itsi	3.64
3.64	Pron	ʃini↗↓	4.00
3.64	Orth	xiní	4.00
4.00	Pron	raˈ	4.32
4.00	Orth	ra	4.32
4.61	Tokens	3	6.13
5.07	Eng	hat	5.89
5.07	Orth	xini	5.89
5.07	Esp	sombrero	5.89
5.07	Pron	ʃiɻni↗	5.89
6.13	Tokens	4	9.61
6.40	Pron	inn̪ii↘↗	7.35
6.40	Orth	itiin	7.35
6.40	Eng	An insect was stuck on his hat.	9.28
6.40	Esp	Se pegó un insecto en el sombrero de él.	9.28
7.81	Orth	itsi	8.36
7.81	Pron	iˈtsiˈ	8.36
8.36	Pron	ʃiɻni↗↓	8.96
8.36	Orth	xini	8.96

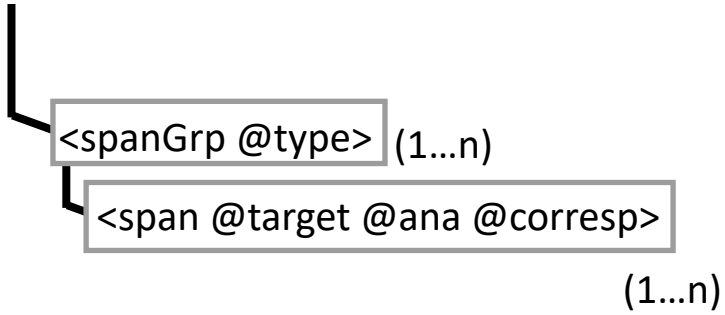


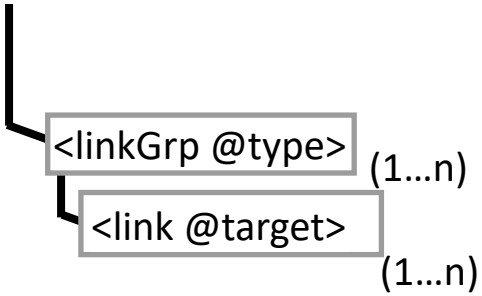
TEI Output from Praat

```
<timeline>
....
</timeline>
 <annotationBlock>
      …
   <u n="4" xml:id="d56e0" start="6.13" end="9.61">
     <seg function="utterance"  notation="orth" xml:id="T-seg-orth-6.13" sameAs="#T-seg-pron-6.13" >
       <w synch="#T6.40"  xml:id="T-orth6.40">itiin</w>
       <w synch ="#T7.81"  xml:id ="T-orth7.81">itsi</w>
       <w synch ="#T8.36"  xml:id ="T-orth8.36">xini</w>
       <w synch ="#T8.97"  xml:id ="T-orth8.97">ra</w>
     </seg>
     <seg function="utterance" notation="ipa"  xml:id="T-seg-pron-6.13" sameAs="#T-seg-orth-6.13">
       <w synch="#T6.40"  xml:id="T-pron6.40">inʈii↘↗</w>
       <w synch="#T7.81"  xml:id="T-pron7.81">iɭtsiɭ</w>
       <w synch="#T8.36"  xml:id="T-pron8.36">ʃiɻni↗↓</w>
       <w synch="#T8.97"  xml:id="T-pron8.97">ɾaɭ</w>
     </seg>
   </u>
   <spanGrp type="translation">
     <span xml:lang="es" target="#T-seg-orth-6.13 #T-seg-pron-6.13">Se pegó un insecto en el sombrero de él.</span>
     <span xml:lang="en" target="#T-seg-orth-6.13 #T-seg-pron-6.13">An insect sat on his hat.</span>
   </spanGrp>
 </annotationBlock>
```

```
<timeline>
  …..
  <when xml:id="T6.40" interval ="6.40"/>
  <when xml:id ="T7.35" interval ="7.35"/>
  <when xml:id="T7.81" interval ="7.81"/>
  <when xml:id ="T8.36" interval ="8.36"/>
  <when xml:id="T8.97" interval ="8.97"/>
  <when xml:id="T9.28" interval ="9.28"/>
</timeline>
```

Standoff Annotation in TEI: <spanGrp> & <linkGrp>:





<spanGrp> is used to annotate: Translations *(English, Spanish),* grammar, Semantics (multiple aspects), Interlinear glossed text, General editorial notes

<linkGrp> links (via <link @target>) pre-existing translation content

- Point to separated language content (usually <w> or <seg>)
- Requires @xml:id for all values to be annotated
- Can be included in most TEI be inserted close to target content
- Structure and tag content correspond to feature structure inventory <fs>

SIL Documents: Annotating translations

```
<item>
  <graphic url="Aves-02.png"/>
  <w xml:id="d1e53" xml:lang="mix">
    <w xml:id="d1e54">chumi</w>
    <w xml:id="d1e56">xini</w>
    <w xml:id="d1e58">ka'nu</w>
  </w>
  <w xml:id="d1e60" xml:lang="es-MEX">
    <w xml:id="d1e61">tecolote</w>
  </w>
  <w xml:id="d1e63" xml:lang="es">
    <w xml:id="d1e64">búho</w>
    <w xml:id="d1e66">cornado</w>
  </w>
</item>
```

Annotations: Translations

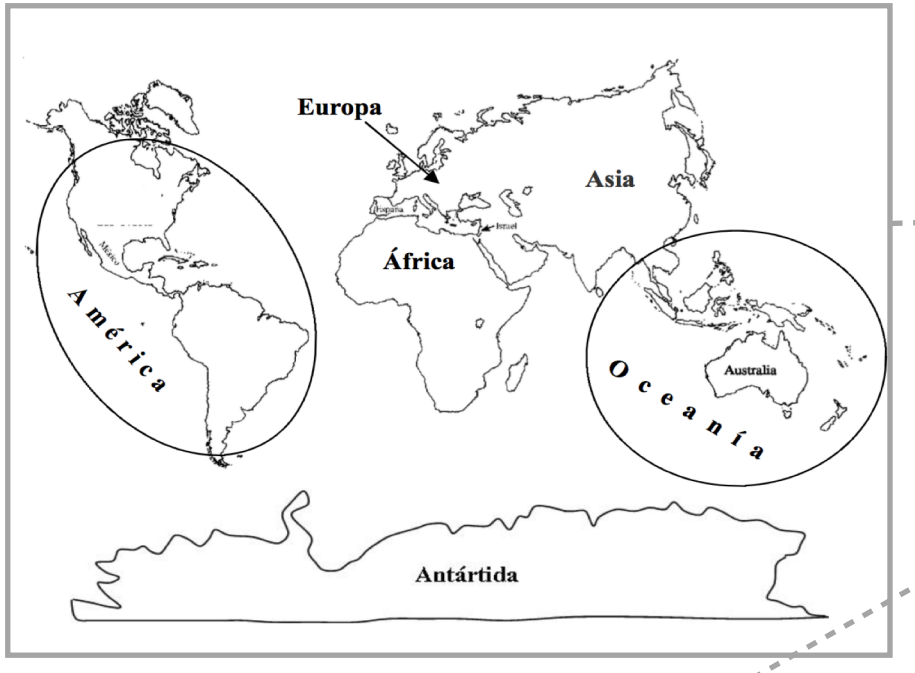
```
<linkGrp type="translation">
  <link target="#d1e53 #d1e60"/>
  <link target="#d1e53 #d1e63"/>
</linkGrp>
```

```
<spanGrp type="translation">
  <span xml:lang="en" target="#d1e53">Great Horned Owl</span>
  <span xml:lang="la" target="#d1e53">Bubo virginianus</span>
</spanGrp>
```

SIL Documents: Basic Prose markup

PDF Source

Ñu'u Ncha'i ka



Yee ñu'u tsi chikuii nuu Ñu'u Ncha'i. Yee kua'a ka chikuii cha xoo ka ñu'u. Yee ñu'u luu ka nania "islas", cha inkai ma'i chikuii. Cha ñu'u ka'nu ka nania "continente". Yee iñu "continente" nania: África, América, Antártida, Asia, Europa tsi Oceanía .

TEI

```
<div xml:id="L145-13">
  <head>
    <seg xml:id="L145-13-00" type="subject">
      <w xml:id="d1e1437">
        <w xml:id="d1e1438">Ñu'u</w>
        <w xml:id="d1e1441">Ncha'i</w>
      </w>
      <w xml:id="d1e1444">ka</w>
    </seg>
  </head>
  <head><graphic url="L145_10.jpeg"/></head>
  <p>
    <seg xml:id="L145-13-01" type="S">
      <w xml:id="d1e1458">Yee</w>
      <w xml:id="d1e1461">ñu'u</w>
      <w xml:id="d1e1464">tsi</w>
      <w xml:id="d1e1467">chikuii</w>
      <w xml:id="d1e1470">nuu</w>
      <w xml:id="d1e1471">
        <w xml:id="d1e1473">Ñu'u</w>
        <w xml:id="d1e1477">Ncha'i</w>
      </w>
      <pc>.</pc>
    </seg>
          ....
</div>
```

SIL Documents: Prose annotation

```
<div xml:id="L145-13">
  …
 <seg xml:id="L145-13-01" type="S">
  <w xml:id="d1e1458">Yee</w>
  <w xml:id="d1e1461">ñu'u</w>
  <w xml:id="d1e1464">tsi</w>
  <w xml:id="d1e1467">chikuii</w>
  <w xml:id="d1e1470">nuu</w>
  <w xml:id="d1e1471">
    <w xml:id="d1e1473">Ñu'u</w>
    <w xml:id="d1e1477">Ncha'i</w>
  </w>
  <pc>.</pc>
 </seg>
  …
</div>
```

Annotations: Translations

```
<spanGrp type="translation">
 <span target="#L145-13-01"xml:lang="en">There is land and water on the Earth.</span>
 <span target="#L145-13-01" xml:lang="es">Hay tierra y agua en la Tierra.</span>

 <span target="#d1e1458" xml:lang="en">there is</span>
 <span target="#d1e1458" xml:lang="es">hay</span>

 <span target="#d1e1461" xml:lang="en">land</span>
 <span target="#d1e1461" xml:lang="es">tierra</span>

 <span target="#d1e1464" xml:lang="en">and</span>
 <span target="#d1e1464" xml:lang="es">y</span>

 <span target="#d1e1467" xml:lang="en">water</span>
 <span target="#d1e1467" xml:lang="es">agua</span>

 <span target="#d1e1471" xml:lang="en">Earth</span>
 <span target="#d1e1471" xml:lang="es">tierra</span>
</spanGrp>
```

SIL Documents: Prose annotation

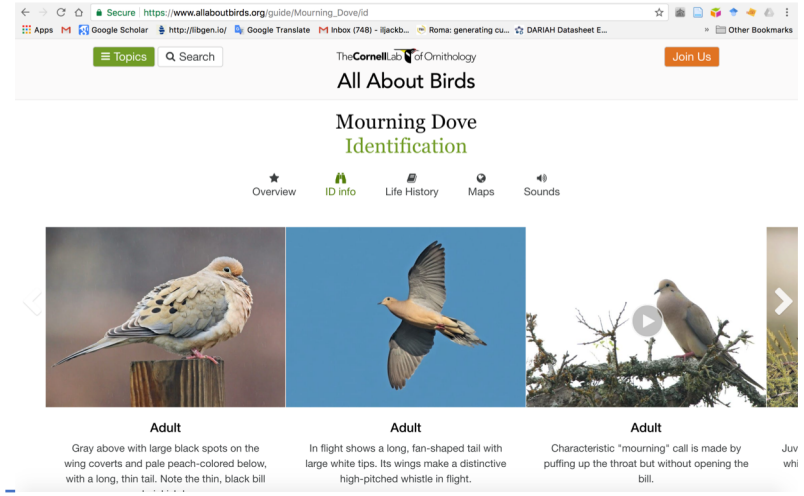
```
<div xml:id="L145-13">
  …
 <seg xml:id="L145-13-01" type="S">
  <w xml:id="d1e1458">Yee</w>
  <w xml:id="d1e1461">ñu'u</w>
  <w xml:id="d1e1464">tsi</w>
  <w xml:id="d1e1467">chikuii</w>
  <w xml:id="d1e1470">nuu</w>
  <w xml:id="d1e1471">
    <w xml:id="d1e1473">Ñu'u</w>
    <w xml:id="d1e1477">Ncha'i</w>
  </w>
  <pc>.</pc>
 </seg>
  …
</div>
```

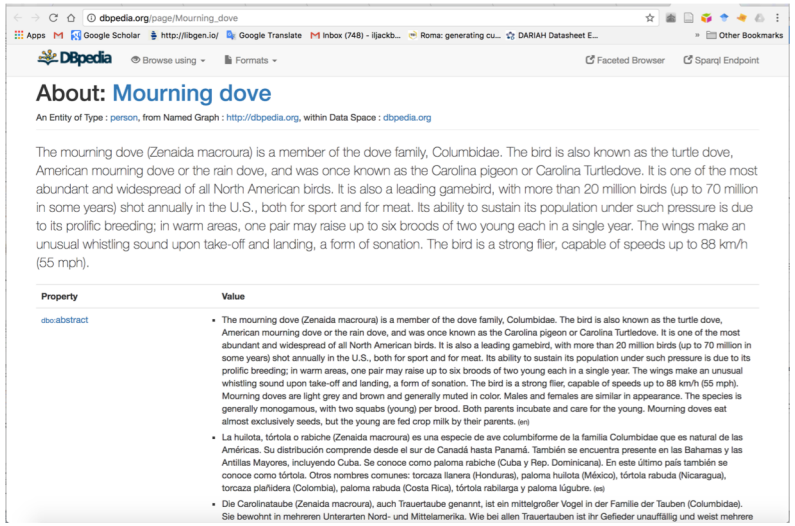
Annotations: Grammar

```
<spanGrp type="gram">
  <span target="#d1e1458" type="pos" ana="#COP-EXIST"/>
  <span target="#d1e1458" type= "mood" ana="#REAL"/>
  <span target="#d1e1461" type="pos" ana="#N"/>
  <span target="#d1e1464" type="pos" ana="CONJ-COORD"/>
  <span target="#d1e1467" type="pos" ana="#N"/>
  <span target="#d1e1471" type="pos" ana="#ADPOS"/>
  <span target="#d1e1471" type="pos" ana="#N"/>
</spanGrp>
```

TEI Dictionary: (data view)

```
<entry xml:id="bird-mourning_dove">
  <form type="lemma">
    <orth xml:lang="mix">lakuku</orth>
    <pron xml:lang="mix" notation="ipa" cert="medium">la-ku-kuˀ</pron>
    <ptr type="sound" target="/media/N_mourning_dove.wav"/>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
  </form>
  <sense corresp="http://dbpedia.org/resource/Mourning_dove">
    <usg type="domain" corresp="http://dbpedia.org/resource/Animal">Animal</usg>
    <xr type="hyponymOf">
      <ref corresp="#bird" xml:lang="en">bird</ref>
      <ref corresp="#bird" xml:lang="mix">saa</ref>
      <ref  type="sense" corresp="http://dbpedia.org/resource/Bird"/>
    </xr>
    <cit type="translation">
      <form><orth xml:lang="en">mourning dove</orth></form>
    </cit>
    <cit type="translation">
      <form><orth xml:lang="es">tortolita</orth></form>
    </cit>
  </sense>
  <cit type="example" corresp="/SIL_docs/L152/L152-tok.xml#L152-01-01">
    <quote xml:lang="mix">In kii ra in <oRef>lakuku</oRef> kunia tanta'i tsi in ncho'o, cha koo xu'in sa'i viko.</quote>
  </cit>
  <!-- could also include references to images (where available) -->
</entry>
```





TEI Dictionary: (User view)

Lakuku (noun)
[la˩ku˥ku˥]

(Animal)
type of Bird
in Saa

En. mourning dove
Es. tortolita

Example:
In kii ra in ***lakuku*** kunia tanta'i tsi in ncho'o, cha koo xu'in sa'i viko.
En. One day a mourning dove wanted to get married to a hummingbird, but didn't have the money for the party.
Es. Un día una tortolita quería casarse con un colibrí, pero no tenia dinero para la boda.

Bowers, J., & Romary, L. (2018). Bridging the gaps between digital humanities, lexicography and linguistics: a TEI dictionary for the documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America*, *39*(2).

Editing and Normalizing Speaker Authored Sources:

Source written by speaker

Tsika *kuan *yuu *ka *un treen ka

kua'an yu kaa nuu

TEI Encoding

```
<seg xml:id="d1e13171" xml:lang="mix" resp="#TS" type="S">
  <w xml:id="d1e13172">Tsika</w>
  <w xml:id="d1e13174" orig="kuan">kua'an</w>
  <w xml:id="d1e13176" orig="yuu">yu</w>
  <w xml:id="d1e13178" cert="high" orig="ka">kaa</w>
  <w xml:id="d1e13180" orig="un">nuu</w>
  <w xml:id="d1e13182">treen</w>
  <w xml:id="d1e13184">ka</w>
  <pc>.</pc>
</seg>
```

- Tag speaker/author of MIX content using @resp
- Correct/normalize spelling, preserve original with @orig
- (also @norm available to keep original and stor normalized form in attribute)
- Editorial certainty indicated with @cert

Integrating Content from Academic Papers: Pike & Ibach (1978)

Paper is the original benchmark for the language's tonal system

Contains a significant amount of vocabulary examples and their tones

However a lack of standardization in both their phonetic alphabet and their tone characterization creates an enormous amount of work to normalize and integrate into project's data model (IPA & TEI)

PDF source scanned, OCR works on English but not Mixtec

Doesn't use normal standards of phonetic representation!

Non-IPA

ȼ = t͡s
š = ʃ
č = t͡ʃ
ẓ = t͡z
ǰ = d͡ʒ
g̱ = k̬

Tone levels are reversed from conventional description

^{3} = Low → ˩
^{2} = Mid → ˧
^{1} = High → ˥

Strings are often interrupted

š[i.]^{2}ši^{3}-ȼi^{2} '

k[o.]^{1}lo^{1}

Integrating Content from Academic Papers: Pike & Ibach (1978)

1. PHONOLOGICAL WORD

In Mixtepec Mixtec[1] the minimal phonological word is made up of the sequence of two syllables – a couplet. This couplet is the complex nucleus of the word; the first syllable is marked phonologically by a lengthened vowel unless that vowel is preceding /ʔ/: *k[o.]¹lo¹ko¹* 'our (excl.) male turkey', *š[i.]²ši³-ɛi²* 'his or her (child) aunt', *s[o.]³ko³-yu³* 'my (polite)

Source pdf (scanned)

TEI Encoding

```
<div>
  <label>1. PHONOLOGICAL WORD</label>
  <p> ….is preceding /<c notation="ipa">ʔ</c>/:

    <seg xml:id="d1e24" xml:lang="mix" notation="ipa">
      <w xml:id="d1e25" orig="k[o.]¹lo¹">koˀloˀ</w> <w xml:id="d1e27" orig="ko¹">koˀ</w>
    </seg>
    '<seg xml:id="d1e28" xml:lang="en" notation="orth">
      <w  xml:id="d1e29a">our</w> <w  xml:id="d1e29b">(excl.)</w>
      <w xml:id="d1e30a">male</w> <w xml:id="d1e30b">turkey</w></seg>',

    <linkGrp type="translation">
      <link target="#d1e24 #d1e28"/>
    </linkGrp>
    ……..
  </p>
……..
</div>
```

Challenges in integrating diverse resources

- “Lone wolf” & data logjam
- Adding new diverse data as found requires time and new workflows
- Where to stop!?
- Multi-layered linguistic annotations and translations time consuming;
- Normalization of extensive variation in orthography, and phonetic transcriptions from external sources;
- Moving targets (changing orthography) & as of last yr+ nobody to make suggestions to regarding orthography
- Use of emerging technology no automatic tools, and ever increasing diversity of data structure input makes limits capacity for automation and make manual treatment necessary

Key Conclusions

- TEI has the capacity to handle needs of LD data, representation;
- Applying TEI to LD and indigenous language strengthens the standard
- There is a need for existing tools (Flex etc.) to enable TEI output (currently only EXMARaLDA allows TEI output)
- LD projects would benefit from adopting TEI as working or output data format in the interest of compatibility, interchange and stability
- Time needed to process and deal with all of the aforementioned issues makes my work less user friendly for the community

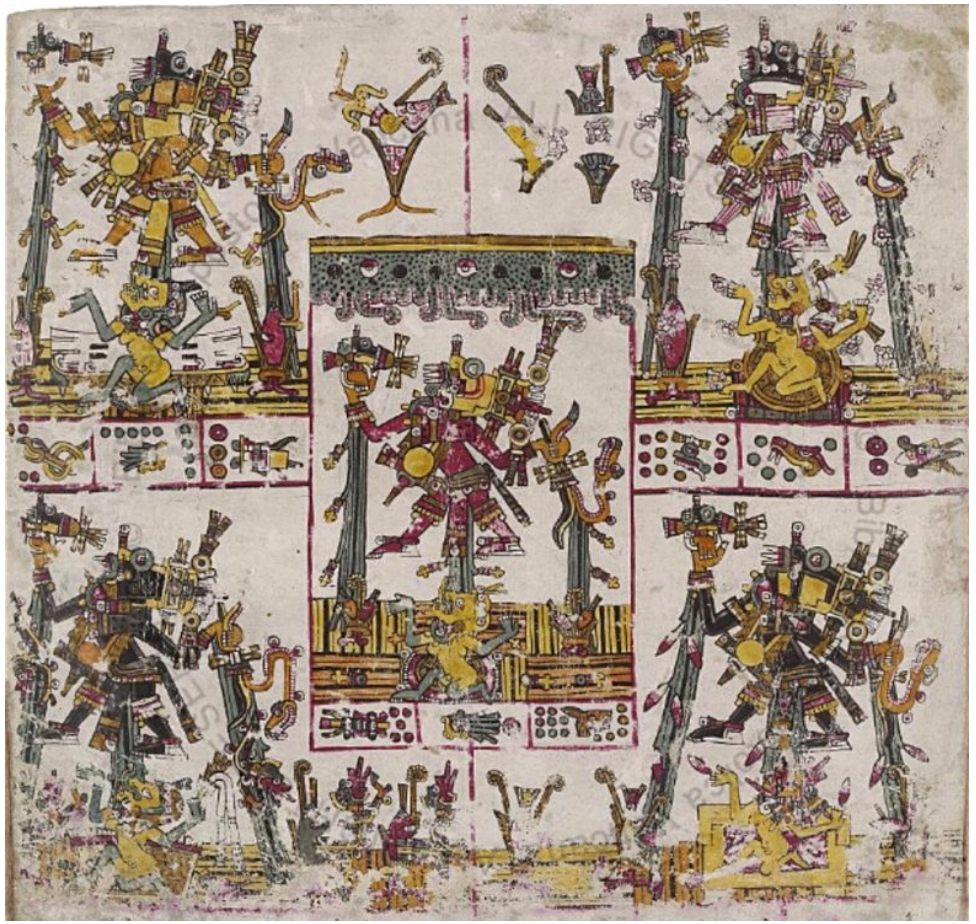
Moving Forward

- Need to make data user friendly, and accessible to community (online portal)
- Make reusable for the creation of learning materials
- Find non-specialist tools for editing the data
- Pursue grants to continue project, hire speakers as editors and contributors
- Utilize phonetic material to train ML to automate annotation
- Bootstrap one translation language to produce the other (where only one is present)
- Automate collection of conceptual data (for annotations) w/ webscraping, other methods
- Need for documentation projects to have long-term view in the data format outputs (standards)
- TEI stabilization of annotations will give rise to tools to ease encoding process

Mahalo! Tatsa’vi kueni!

Special thanks to my consultants Geremaia and Tisu’ma Salazar for sharing their language with me!

iljackb@gmail.com
https://github.com/iljackb/Mixtepec_Mixtec



Codex Borgia, c. 1500, p. 28 (Vatican Library); obtained from:
Dr. Helen Burgos Ellis, "Codex Borgia," in *Smarthistory*, September 11, 2017, accessed March 2, 2019, **https://smarthistory.org/codex-borgia/**.